Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 5 | Issue : 11 | Nov 2019

EMAIL CLASSIFICATIONS FOR SPAM MAIL DETECTION BY COMPARING THREE DIFFERENT ALGORITHMS USING WEKA

Vutharkar Nagaveni [1] | Dr. Vimal Pandya [2]

[1] Computer Science, Rai University, Saroda, Gujrat, India.

[2] Director, Navgujarat College Of Computer Applications, Ashram Road, Ahmedabad, Gujrat, India.

ABSTRACT

Email is the system for sending messages from one individual to another via telecommunications links between computers or terminals using dedicated software applications. Nowadays, Email is used most common and effective mode of communication way to communicate in personal, individual and professional level. As increase of email users there will be increase of spam emails from the past few years. This paper explore how email data was classified using three different classifiers (Naive Bayes classifier ,Support Vector Classifier,J48 Classifier) for detecting spam using WEKA. This experiment was performed based on dataset to find spam in different parameters like finding Accuracy, Recall, Precision, Fmeasures and False Position Rate etc. The final classification result should be '1' if it is finally spam present , otherwise, it should be '0' for no spam. Finally this paper shows that J48 classifier is best and efficient algorithm for detection of spam emails for dataset that classified as binary tree among other algorithms.

KEY WORDS: Decision tree, WEKA, dataset, classification algorithm,,Support vector Machine, Email.

I. INTRODUCTION:

Spam is commonly defined as unsolicited bulk email messages received without one's permissions, and the goal of spam detection is to distinguish between spam and legitimate email messages. Most of the spam contain viruses, Trojan horses and other harmful software that may cause to failures computer systems, networks, bandwidth and storage space to slows down email servers.

Now a days, spam mails have been increased alarmingly, prompting a need for anti-spam filters which are reliable, accurate, and can effectively classify legitimate mails from spam. There are several text mining and machine-learning techniques to classify spam mail have been used such as Naïve Bayes, and Support Vector Machines, J48 Classifier, Bayes net classification and etc.

Spammers collects email IDs from various sources such as chats, websites, newsgroups , malware and address details of users, which are easly available from other spammers for low price and bulk of messages are sent to receipients where, the volumes of which create enormous productivity losses to IT firms and huge serious security threats that carriers classified information.
Hence, the classification of emails is prime importance to handle spam emails.

Machine learning algorithms are used for classification of objects in different classes to prove efficient in classifying emails as spam or harm. In my research work I, used three main machine learning algorithms namely, Naive Bayes classifier, Support Vector Classifier and J48 Classifier for classification.

WEKA, which is a free, open-source software that compiles data-mining algorithms for machine-learning applications. WEKA is capable of performing tasks such as pre-processing, statistical processing and visualization of data (www.cs.waikato.ac.nz/ml/weka). Algorithms such as Naive Bayes classifier, Support Vector Classifier and J48 Classifier are applied in classifying spam mail were presented. Descriptions of these algorithms and a comparison of their performance using the WEKA environment have been reported.

We have given a short review detailed description of the three classification algorithms and present the experimental details followed by results and discussion. Finally, we present the conclusions followed by avenues for future work.

II. Related Work: In this study contains various previous work done by researchers for classification of spam emails in brief is presented. Data-mining classification algorithms have been deployed to filter the spam and the legitimate emails. A study was conducted to compare four algorithms (J48, ID3, Alternating Decision Tree, and Simple CART) for classification accuracy Spam datasets were run through the algorithms in a WEKA environment and it was seen that J48 outperformed other than three.

Awad (2011) conducted a review to assess well-known machine-learning tools (k-NN, Bayesian classification, SVMs, Artificial Immune System, ANNs, and Rough Sets) which suits for classifying spam emails. Where compared using the Spam Assassin spam corpus and confirmed that Naïve Bayes and Rough Set methods showed promise. Hence it is required that Naïve Bayes and Artificial Immune System to improve the performance of hybrid systems, by sorting the dependence of features in Naïve Bayes classifier or by a hybrid of the immune system by Rough Sets. In which Hybrid systems show most promise in spam-filtering efficiency.

In another study which is conducted using the TANGARA data-mining tool to identify the efficient spam classifiers (Kumar et al., 2012). At First construction and selection were carried out on the dataset. Next, the algorithms are used on dataset, and after cross validation took place and the best one chosen based on the error rate, precision and recall. It is observed that Rnd Tree algorithm was the most superior one with 99% accuracy more than other.

In another study, comparison was made among five classification algorithms, namely, Simple Cart, ADTree, J48, Naive Bays, ZeroR, and Random Forest Classification Algorithm to assess ability and select the course, which best suited for the students based on choices (Aher and Lobo, 2012). WEKA, is used to describe and evaluate the result. It was observed that ADTree algorithm is a better choice for wrongly classified instances which are less than the other algorithms.

In (Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza, 2014), Authors have been evaluated various ensemble classifiers for spammer detection in social network. The dataset was taken from Facebook in which spammer behavior has been injected by author. Instead of using content based features, new network structure based features were used to detect the spammers. Some of base classifiers namely J48, IBK, and Naïve Bayes available in WEKA were used and evaluated. Ensemble learning approach of bagging and boosting were used on base classifiers (J48, IBK and Naïve Bayes) and evaluated by using given dataset. In this Bagging ensemble learning approach using J48 is performed well and better than other evaluated classifiers.

In (Trivedi, Shrawan Kumar, and Shubhamoy Dey,2013), Authors were compared the performance of probabilistic classifiers with and without the help of various boosting algorithm. The dataset was taken from Enron email dataset. Genetic Search algorithm was used to select important features, in which 134 features were selected out of 1359 features. Naïve bayes and Bayesian classifiers were evaluated first after that boosting algorithms is used to enhance the performance of the classifiers. Bayesian classifier is performed better than Naïve bayes. Boosting with Resample using Bayesian Classifier has given best result with an accuracy of 92.9% than other classifiers. Adaboost has also given better results. As future work, boosting algorithms can be used with other base classifiers to do the comparison on performance.

Various machine-learning algorithms for the filtration of spam were discussed and compared in another study (Islam and Chowdhury, 2005). This study covered automated filtration and machine-learning methods based on rules and content and used individualized, vector machines for collaborative support, and algorithms that were kernel-based for checking spam. All such techniques were compared with the advantages and presented.

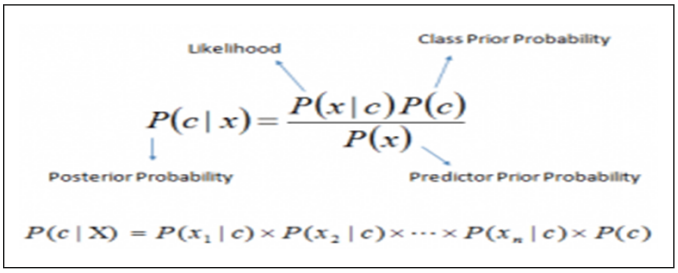
tamaPei-yu et al. (2009) Author suggested an improved Bayesian algorithm approach for classifying spam in which accuracy and simplicity of the KNN algorithm were filters spams using k-nearest neighbor and used for spam filteration. SVM is also used for filtering spam and finds hyper plane to classify the legitimate and spam mails which works with smaller training set.

III. DESCRIPTION OF CLASSIFICATION ALGORITHMS:

A. Naive Bayes Classification Algorithm: It is a classification technique based on Bayes' Theorem which is an assumes independence of among predictors. Naive Bayes classifier states that "presence of one particular feature in a class is unrelated to the presence of any other feature classes".

Naive Bayes model is easy to build and useful for big datasets, which is outperform for highly sophisticated classification methods.

Bayes theorem provides a way to calculate posterior probability P(c|x) from P(c), P(x) and P(x|c) and equation is as below:



In above equation,

- P(c|x) is the posterior probability of class (c,target) given predictor (x,attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Now suppose the event C represents a spam and X is 'containing certain words'. Bayesian filtering would predict the probability that the message is really spam, given the 'test results' (which are certain words) Improved J48 Classification Algorithm for the Prediction of Diabetes on performes of condition.

Naive Bayes algorithm works: The steps to perform the Naive Bayes algorithm is as follows.

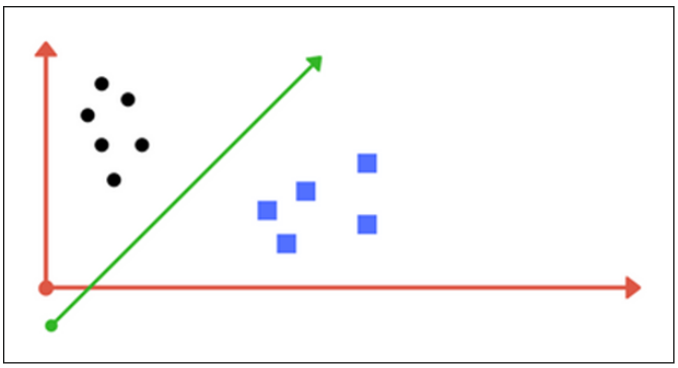
Step 1: Convert the data set into a frequency table.

Step 2: Create Likelihood table by finding the probabilities.

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Naive Bayes works with lesser training set of data to assess the classification parameters, that contains few advantages.

B. Support Vector Machine Classifier Algorithm: Support Vector Machine (SVM) is a discriminative classifier which is defined by a separating hyperplane. In two dimentional space, hyperplane is a line which divides a plane in two parts where each class lay inside either outside of the line.



Support vector machine algorithm works: Process for classifying whether an email is a spam or ham.

1) Collect the dataset
2) Filter the collected data
3) Separate all the messages into tokens codes
4) Make vector with the token code and its appearance frequency calculation

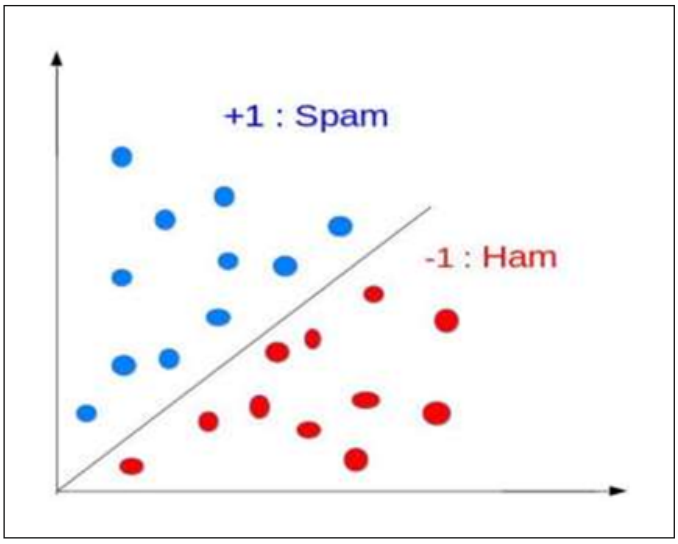
   (X1; Y1), (X2; Y2).... (Xn; Yn)

   where

   Xi is a vector with a numeric value as the number of times token occurs in the message.

   Yi (+1, -1)

   which define two classes, +1 = Spam, -1 = Ham
5) Finally SVM constructs hyper plane by plotting a vector point's to the class +1 for spam class -1 for ham.
6) Classify the data as spam or ham.



C. J48 Classifier algorithm: Decision Tree Algorithm is used to find out the way that attributes-vector behaves for a number of instances. This algorithm generates the rules for the prediction of the target variable and helps to understand the critical distribution of data easliy.

J48 is an extension of ID3. This additional features enhance accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning.

Decision trees are built using J48 with input set (S=s1, s2, s3,...) of training data by using the concept of information entropy. Training data is sample classified by every sample si, which consist as a p-dimensional vector (x1i, x2i, x3i,...), where xj stands for features or attribute values of the sample and where the class si belongs to that.

At each node, the algorithm selects the data attribute, which most adequately divide its sample sets into subsets, which are enriched in any one of the classes. The division condition is the gain of normalized information (entropy difference). This characteristic have the greatest gain of normalized information ,which selected for a decision purpose. This repeats for the smaller sub-lists too.

J48 Classifier algorithm works:
Base cases for this algorithm is as follows.

1. A list of samples reside in the same class, which occurrence forms a leaf node in the decision tree that enforces it to select the class.
2. When no characteristics contribute to any information gain, a decision node is formed a higher level in the tree and utilize the class expected value.
3. When a class not seen earlier turns up, a decision node higher up is formed and utilize the class expected value.

IV. EXPERIMENTS SETUP:

In this section, a report on experimental evaluation of the three algorithms is presented. Firstly the dataset containing the spam email used for evaluation purposes is described in details. After then a short description of the performance

measures of formulas is outlined. Finally, the experimental results and related discussions details are presented in the form of bar chart and table format.

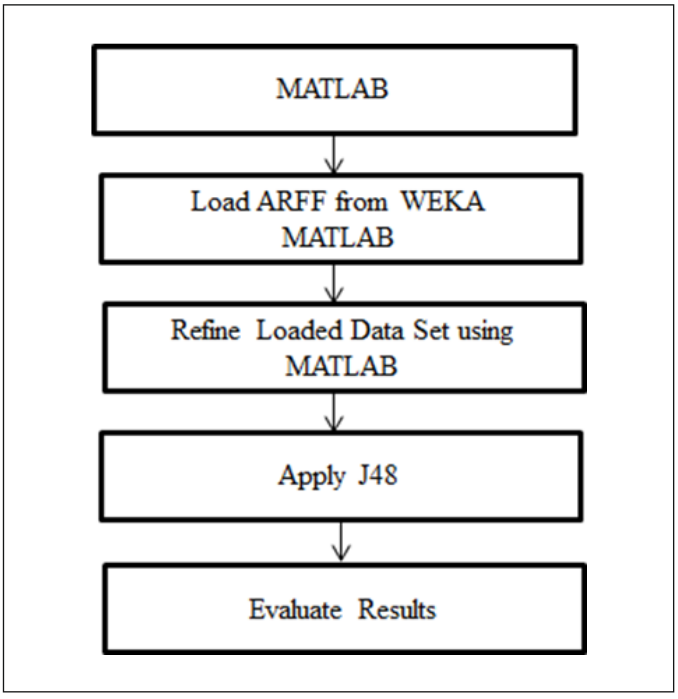
a. WEKA: The University of Waikato in New Zealand developed WEKA (Waikato Environment for Knowledge Analysis),Which is an innovation tool in data mining and machine learning research communities enviorment.This tool was developed by WEKA team since 1994 and contains many inbuilt algorithms for data mining and machine learning systems. It is open source and user friendly with freely available platform-independent software machine learning system. User who are not familiar and does't have knowledge about data mining can also use this software very easily ,as it provides flexible facilities for scripting and experiments evaluations. As on new algorithms appear in research literature are updated in software and avail for users.

b. Steps: The steps to performe using data mining in WEKA is as follows:

- Data pre-processing and visualization
- Attribute selection
- Classification (Decision trees)
- Prediction (Nearest neighbour)
- Model evaluation
- Clustering (Cobweb, K-means)
- Association rules

The algorithms uses WEKA as API in MATLAB. WEKA is a comprehensive open source Machine Learning system toolkit, written in Java platform. The basic functions provide MATLAB interface to WEKA for allow the transformation of data back and forth to use features available in WEKA such as training classifiers.

The algorithm is evaluated by loading arff data file into METLAB from WEKA. After then refining the dataset is done by applying classifier.Finally the results are obtained ,which shows the accuracy and error rate etc. The above flow chart show the algorithms evalution process.



c. Experimental Datasets: This study contains a spam email dataset which is available publicly from the UCI Machine Learning Repository, i.e.SPAM E-mail Database. Which contains 57 attributes and 4601 emails, with 1813 emails being spam and the rest (2788) being normal emails. The dataset is multivariate with real integer attributes and values.

d. Performance Evaluation: This section show the comparison of the different data mining algorithms. In machine learning, particularly for statistical classification, a explanatory table is prepared that permits us to visualize the performance of the algorithm. This is the confusion or error matrix. Its columns represent the predicted class entries and rows represent those of the actual class (Table 1). It is called so because this matrix makes it simple to check if the system is incorrectly identifying two classes (e.g. reverse labelling errors).

Table 1. Confusion Matrix

		Predicted	
		Belongs	Does not Belong
Actual	Belongs	True Positive (TP)	False Negative (FN)
	Does not belong	False Positive (FP)	True Negative (TN)

We then define the following parameters:

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \qquad Recall\,(R) = \frac{TP}{TP + FN}$$

$$Precision\,(P) = \frac{TP}{FP + TP} \qquad False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$Fmeasure = \frac{2 \times P \times R}{P + R}$$

The performance of the algorithms was analised on the basis of a comparison of the above four parameters.

V. RESULTS AND DISCUSSION:

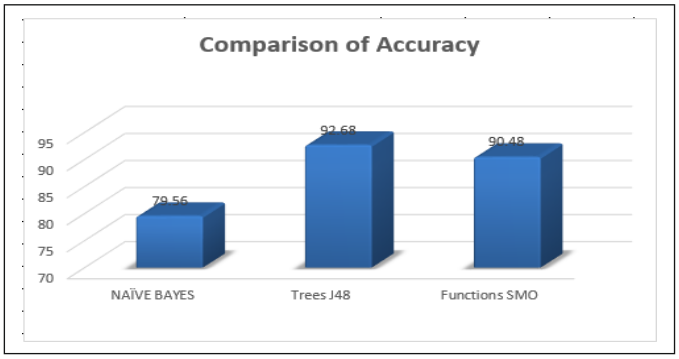
The spam email dataset was used as inputs to the algorithms which were run in the WEKA environment and final results are shown as follows.

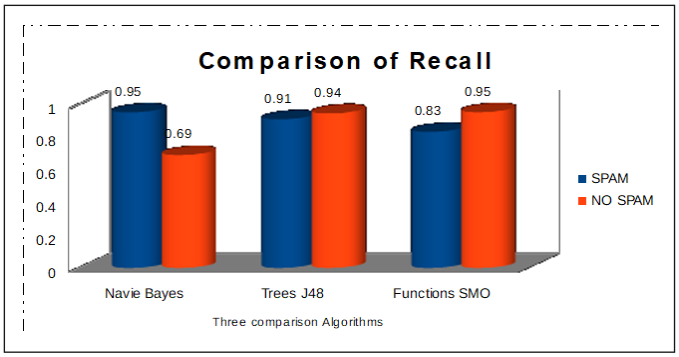
The results are divided in 2 categories in the class coloumn: Spam as 1 and No Spam as 0 (Legitimate).

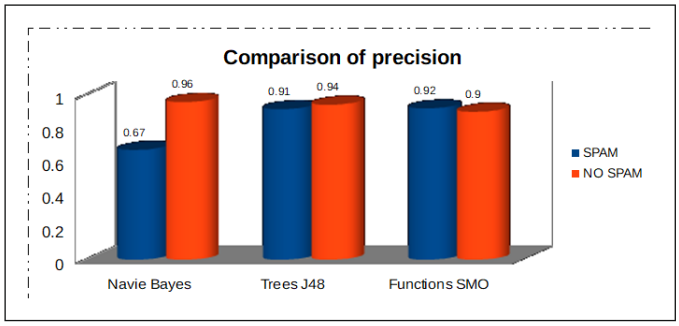
Table 2: A comparison of results between three algorithms

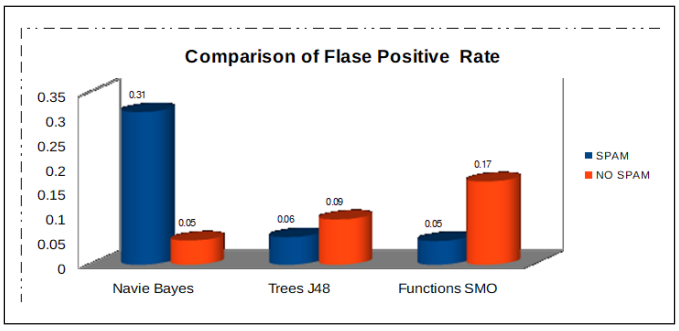
Classification algorithm	Accuracy (TP+TN)/ (TP+TN+FP+FN)	Recall TP/(TP+FN)	Precision TP/(TP+FP)	FP RATE FP/(FP+TN)	TP RATE (=RECALL)	F MEASURE 2*PR/(P+R
NAÏVE BAYES	79.56	0.951	0.666	0.310	0.951	0.784
		0.690	0.956	0.049	0.690	0.801
J48	92.68	0.908	0.913	0.056	0.908	0.911
		0.944	0.940	0.092	0.944	0.942
SVM	90.48	0.831	0.918	0.048	0.831	0.873
		0.952	0.897	0.169	0.952	0.923

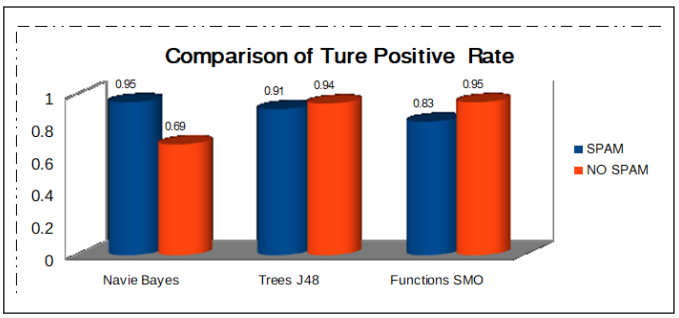
Classifcation Algorithm	Confusion Matrix		
	a	b	Classifèd as
NAIVE BAYES	1725	88	a=1 b=0
	865	1923	
J48	1646	167	a=1 b=0
	156	2632	
Functions SMO	1507	306	a=1 b=0
	134	2654	

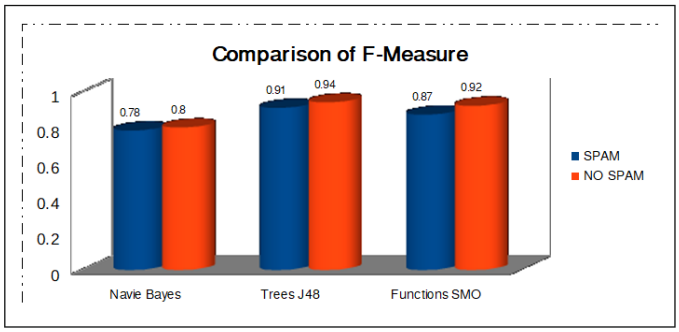












Features	Ranking			Categories
	1	2	3	
Accuracy	J48	NAÏVE BAYES	SVM	
Precision	SVM	J48	NAÏVE BAYES	Spam
	NAÏVE BAYES	J48	SVM	No Spam
Recall/ TPR	NAÏVE BAYES	J48	SVM	Spam
	SVM	J48	NAÏVE BAYES	No Spam
F-Measure	J48	SVM	NAÏVE BAYES	Spam
	J48	SVM	NAÏVE BAYES	No Spam
False-Positive Rate (FPR)	NAÏVE BAYES	J48	SVM	Spam
	SVM	J48	NAÏVE BAYES	No Spam

In above table it indicates that J48 is the best algorithm in terms of accuracy and also performs better in Recall, Precision, FPR and F-measure.

J48 creates decision trees from a labelled data and utilized to take a decision by dividing the data as reduced subsets for the study. The normalized data added information or entropy variation during splitting is done. The decision is taken based on the maximum normalization by gain of attributes to process.

The ability of J48 decision trees is used for missing values, value ranges, etc. Which makes it a superior algorithm compaired to other algorithms. In this study it is observed that no algorithm shows 100% accuracy for finding spam in Email classification .

VI. CONCLUSION AND FUTURE WORK:

This paper introduces a method to classify mails based on three classifiers, i.e. J48, SVM, and Naïve Bayes. These classifiers were evaluated to separate spam from the email dataset by using WEKA. The emails were identified as spam (1) or not spam (0), which reflected the attributes of the dataset of email for spam filtering.

The algorithm was checked against parameters such as Accuracy, Precision, Recall, F-measure and False Positive Rate. The analysis of the results demonstrated clearly that even though J48 is a very simple classifier which uses a decision tree, it gave the most accurate result in the experiment (92.68%). In addition, it also performed well in other parameters, with highly favorable values and coming first in precision no spam category and F-meature spam category. On the other hand, it retains the second position for other parameters. Thus, it can be understood that J48 is the algorithm which is preferable to other algorithms that are compared in this study for the classification of e-mails with the purpose of filtering spams.

SVM lassifier also showed good results with accuracy of 90.48%% and had a better performance results in other parameters too. But SVM is given accuracy of (79.56%) ,which is poor results in comparison to other classification.

In further research study it is required to improve an depth anaysis algorithm like Genetic algorithm, and classification techniques for finding the spam In addition, different algorithms which are not included in WEKA should be added to test and experiments with various feature of attributes selections for comparisions.

REFERENCES:

1. Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza. "Spammer Classification Using Ensemble Methods over Structural Social Network Features." In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02, pp. 454-458. IEEE Computer Society, 2014.
2. Trivedi, Shrawan Kumar, and Shubhamoy Dey. "Interplay between Probabilistic Classifiers and Boosting Algorithms for Detecting Complex Unsolicited Emails."Journal of Advances in Computer Networks 1, no. 2 (2013): 132-136.
3. Sneha Singh,Sandeep Kaur., "IMPROVED SPAMBASE DATASET PREDICTION

USING SVM RBF KERNEL WITH ADAPTIVE BOOST", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308 ,Volume: 04 Issue: 06 | June-2015.

4. Ghada Hammad AL-Rawashdeh Dr.Rabiei Bin Mamat "Comparison of four email classification algorithms using WEKA", International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 2, February 2019.
5. Sunil Ray, " 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R", https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained, Sept 11th, 2017
6. Diksha S. Jawale, Ashwini G. Mahajan,Kalyani R. Shinkar and Vaishnavi V. Katdare "Hybrid spam detection using machine learning",IJARIIT,ISSN: 2454-132X,Impact factor: 4.295,Volume 4, Issue 2,2018
7. Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.
8. Emmanuel Gbenga Dada, "Machine learning for email spam filtering: review, approaches and open research problems", https://doi.org/10.1016/j.heliyon.2019.e01802 Received 3 September 2018; Received in revised form 25 February 2019; Accepted 20 May 2019 ,2405-8440,Published by Elsevier Ltd.
9. https://en.wikipedia.org/wiki/C4.5_algorithm.
10. K.F.Bindhia,Y.Vijayalakshmi, P.Manimegalai and Suvanam Sasidhar Babu, "Classification Using Decision Tree Approach towards Information Retrieval Keywords Techniques and a Data Mining Implementation Using WEKA Data Set", International Journal of Pure and Applied Mathematics Volume 116 No. 22 2017, 19-29 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) url: http://www.ijpam.eu .